A latent variable model for improving inference in trials assessing the effect of dose on toxicity and composite efficacy endpoints

Journal Title
XX(X):3–30

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

James M. S. Wason[1,2] and Shaun R. Seaman[2]

Prepared using sagej.cls

Abstract

It is often of interest to explore how dose affects the toxicity and efficacy properties of a novel treatment. In oncology, efficacy is often assessed through response, which is defined by a patient having no new tumour lesions and their tumour size shrinking by 30%. Usually response and toxicity are analysed as binary outcomes in early phase trials. Methods have been proposed to improve the efficiency of analysing response by utilising the continuous tumour size information instead of dichotomising it. However these methods do not allow for toxicity or for different doses. Motivated by a phase II trial testing multiple doses of a treatment against placebo, we propose a latent variable model that can estimate the probability of response and no toxicity (or other related outcomes) for different doses. We assess the confidence interval coverage and efficiency properties of the method, compared to methods that do not use the continuous tumour size, in a simulation study and the real study. The coverage is close to nominal when model assumptions are met, although can be below nominal when the model is misspecified. Compared to methods that treat response as binary, the method has confidence intervals with 30-50% narrower widths. The method adds considerable efficiency but care must be taken that the model assumptions are reasonable.

Keywords

Augmented binary method; Composite endpoints; Efficacy/toxicity; Phase I/II.

Introduction

Traditionally phase I trials of a new drug are carried out to identify a safe dose. Once a suitable dose is identified, it is tested for initial signs of efficacy in a phase II trial. A recent trend is that more trials, known as phase I/II, consider how the dose of a treatment affects both efficacy and toxicity at the same time. In oncology, for example, phase I trials are increasingly using dose expansion cohorts to assess for early signs of efficacy [1]; in addition, seamless phase I/II trials have been recommended for improving the efficiency of the drug development process [2].

For trials which have the objective of exploring the effect of dose on both toxicity and efficacy, there are two main statistical questions of interest. The first is the best choice of design to identify a suitable dose in an efficient and ethical way. Many papers have proposed novel adaptive designs that allow this, including [3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15]. A recent book[16] provides an overview of Bayesian methods in this area. The second, related, issue is how to model the data gathered during such a trial at the end. This is important, as the purpose of the trial is to choose the most suitable dose to take forward for a more definitive trial. Many papers that propose adaptive designs also consider suitable models for efficacy and toxicity data. We consider the second question in this paper, motivated by the question of how one could better analyse the probability of early response and no toxicity in an oncology phase II trial that tested two doses of a novel treatment against placebo. Although we primarily consider oncology

[1]Institute of Health and Society, Newcastle University
[2] MRC Biostatistics Unit, University of Cambridge

Corresponding author:
James Wason Institute of Health and Society, Newcastle University, Newcastle upon Tyne, NE2 4BN, UK.
Email: james.wason@newcastle.ac.uk

Prepared using sagej.cls

in this paper, the methods are applicable to any clinical area where efficacy is assessed with a responder-based composite outcome.

Toxicity and efficacy in oncology are usually measured with composite endpoints. In solid tumours, efficacy is evaluated using the RECIST v1.1 criteria [17] (henceforth referred to as RECIST) which categorises patients into complete response, partial response, stable disease and progressive disease according to the change in the sum of longest diameter of target lesions (henceforth simply called tumour size) and other criteria for treatment failure such as whether or not new lesions appear and unacceptable growth in non-target lesions. A standard early measure of efficacy is the response rate, which is the proportion of patients who have a partial or complete response. Toxicity is generally measured using a suitable set of criteria, such as the Common Toxicity Criteria (CTC) [18], which grades different types of adverse events between 1 and 5 depending on severity (with a grade of 3 or above being classed as a severe toxicity). Often toxicity and efficacy are analysed as binary endpoints with suitable models used to account for correlation. Some papers have proposed analysing the variables as ordinal [12, 19, 20, 21]. Others have considered jointly modelling a binary toxicity outcome and a continuous efficacy outcome [6, 22, 23] with the continuous outcome representing a biomarker expression or change in tumour size. In the case of modelling the joint probabilities of efficacy and toxicity, with efficacy defined using the commonly used RECIST criteria, existing methods for jointly modelling a binary toxicity outcome and a continuous efficacy outcome are not sufficient. This is because the RECIST efficacy outcome is not purely continuous but is instead a composite of a continuous outcome (change in tumour size) and binary (new lesions appearing).

In previous work [24] we proposed a method for making inferences on this composite efficacy outcome by fitting a joint model to the continuous and binary efficacy components. This makes inference on the clinically relevant response outcome but considerably improves the efficiency compared to modelling the binary composite outcome alone. The efficiency gain comes from utilising the continuous nature of change in tumour size rather than dichotomising it. To our knowledge, there has been no work on modelling how dose affects the binary toxicity and a dichotomised binary efficacy outcome, whilst utilising continuous information to improve efficiency.

In this paper we develop the method of Wason and Seaman[24] so that it can be used when toxicity is a component of the outcome and there are multiple doses of the same treatment. A multivariate probit model with latent variables for toxicity and new lesion events is used to jointly model the toxicity and the two components of the efficacy outcome. In this paper we concentrate on estimating the probabilities of the four possible combinations of efficacy and toxicity (efficacy and toxicity, efficacy without toxicity, toxicity without efficacy, no efficacy and no toxicity), primarily focusing on the probability of efficacy without toxicity. This quantity is likely to be of considerable interest, as it represents providing patient benefit without unacceptable side-effects. A dose that maximises this quantity will be desirable to take forward to further testing. If other quantities involving response as a binary outcome are of interest, then the method presented here will also provide benefits for estimating them. Following much of the previous work in the area, we consider toxicity as a binary outcome. In the discussion we describe how the method might be extended to allow for more complex toxicity information.

A previously proposed method for phase I/II trials by Zhong et al[10] used a trivariate model for toxicity, efficacy and a more quickly

observed intermediate efficacy marker. Although our method also assumes a trivariate model, both the aim of the method and the model used are different.

As the motivation for the methods in this paper we use the example of the HORIZON II trial [25] (clinicaltrials.gov identifier NCT00384176). HORIZON II is a phase II trial that tested two doses of cediranib (20mg once daily, 30mg once daily) against placebo for the treatment of metastatic colorectal cancer. All patients received infusional fluorouracil, leucovorin, capecitabine and oxaliplatin. We were interested in how one could analyse the effect of dose on the probability of efficacy without toxicity at 6 weeks. This type of outcome is consistent with early phase I/II trials where short-term response outcomes are often used.

The rest of the paper is organised as follows. In Section 2 we define the notation use, and describe the proposed method together with the comparator methods. In Section 3 we describe the simulation study setup and results. In Section 4 we present the analysis of the HORIZON II trial. In Section 5 we summarise the benefits and limitations of the proposed method and discuss potential future work.

Methods

Notation

We consider a set of N patients. The dose (or some suitable transformation of it) given to patient i is denoted by d_i. The ith patient has baseline tumour size y_{i0}, and a single follow-up tumour size measurement of y_{i1}. As in Wason and Seaman[24], we work with the log tumour size ratio, defined as:

$$z_{i1} = \log\left(\frac{y_{i1}}{y_{i0}}\right). \tag{1}$$

This is used because it has been found to be close to normally distributed in real data [26]. Other transformations can be used with minor modification to proposed methods.

In addition to the tumour size data, we also observe whether a patient had an efficacy failure for reasons other than an increase in tumour size (henceforth referred to as non-shrinkage failure). This could be due to new lesions, or an unacceptable increase in the size of non-target lesions. This is summarised by an indicator F_i, which equals 1 if a non-shrinkage failure occurred and 0 otherwise. A separate indicator T_i is equal to 1 if a dose-limiting toxicity occurred and 0 otherwise.

Augmented binary method using a latent variable model

In Wason and Seaman, separate models were fitted for the tumour sizes and for the non-shrinkage failure indicators. The tumour size data were modelled using a multivariate normal model and the non-shrinkage failure data by a series of logistic regression models. Each logistic regression modelled the probability of having a non-shrinkage failure at a certain visit conditional on not having one up to that time and on the tumour size at the beginning of that time period. This allowed correlation between the tumour sizes and non-shrinkage failure indicators to be accounted for, although in a restricted way.

In this paper we propose using an alternative model that allows for correlation between the three components of the efficacy without toxicity outcome. Although one could combine toxicity and non-shrinkage failure and apply the method from Wason and Seaman, this would not model the possibly different relationships between dose and the probabilities of new lesions and toxicity. For example, it is plausible that increasing dose may increase the probability of toxicity but decrease the probability of new lesions. The model we use is

related to the multivariate probit model discussed in, for example, Chib and Greenberg [27].

The model uses log tumour ratio of patient i, Z_{i1} and two latent normally distributed random variables Z_{i2} and Z_{i3}. These are related to F_i and T_i through $F_i = I\{Z_{i2} > 0\}$ and $T_i = I\{Z_{i3} > 0\}$, where $I\{.\}$ represents the indicator function. We focus on a model that allows each component to depend linearly on the dose, but later on more complex relationships are considered. We also allow each component to depend on baseline tumour size. It is plausible that change in tumour size, probability of new lesions and probability of toxicity are each associated with the tumour size at baseline. Thus adjusting for baseline tumour size will improve the precision of other parameter estimates. Other covariates can be adjusted for in a similar manner. The model with linear dependence on dose is:

$$\begin{aligned} Z_{i1} &= \mu_1 + \alpha_1 d_i + \beta_1 y_{i0} + \varepsilon_{i1} \\ Z_{i2} &= \mu_2 + \alpha_2 d_i + \beta_2 y_{i0} + \varepsilon_{i2} \\ Z_{i3} &= \mu_3 + \alpha_3 d_i + \beta_3 y_{i0} + \varepsilon_{i3}, \end{aligned} \tag{2}$$

where the residual error terms have the following distribution:

$$\begin{aligned} (\varepsilon_{i1}, \varepsilon_{i2}, \varepsilon_{i3}) &\sim N(0, \Sigma) \\ \Sigma &= \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1 & \rho_{13}\sigma_1 \\ \rho_{12}\sigma_1 & 1 & \rho_{23} \\ \rho_{13}\sigma_1 & \rho_{23} & 1 \end{pmatrix}. \end{aligned} \tag{3}$$

As is common in multivariate probit models, the error variances of the latent variables are set to 1 for identifiability reasons.

Let $\theta = (\mu_1, \alpha_1, \beta_1, \mu_2, \alpha_2, \beta_2, \mu_3, \alpha_3, \beta_3, \sigma_1^2, \rho_{12}, \rho_{13}, \rho_{23})$. The likelihood contribution for individual i, $l(\theta; Z_{i1}, Z_{i2}, Z_{i3}, d_i, y_{i0})$ can be written as:

$$l(\theta; Z_{i1}, Z_{i2}, Z_{i3}, d_i, y_{i0}) = f(Z_{i1}|d_i, y_{i0}; \theta) f(Z_{i2}, Z_{i3}|Z_{i1}, d_i, y_{i0}; \theta), \quad (4)$$

the product of the marginal pdf for Z_{i1} (conditional on y_{i0} and d_i) and the joint pdf for Z_{i2} and Z_{i3}, conditional on Z_{i1}, y_{i0} and d_i. The marginal pdf of Z_{i1} is normal with mean $\mu_1 + \alpha_1 d_i + \beta_1 y_{i0}$ and variance σ_1^2. The conditional distribution of (Z_{i2}, Z_{i3}) given Z_{i1} is bivariate normal with mean $\tilde{\mu}$ and covariance matrix $\tilde{\Sigma}$, where:

$$\tilde{\mu} = \begin{pmatrix} \mu_2 + \alpha_2 d_i + \beta_2 y_{i0} \\ \mu_3 + \alpha_3 d_i + \beta_3 y_{i0} \end{pmatrix} + \frac{1}{\sigma_1^2} \begin{pmatrix} \rho_{12}\sigma_1 \\ \rho_{13}\sigma_1 \end{pmatrix} (Z_{i1} - \mu_1 - \alpha_1 d_i - \beta_i y_{i0})$$

$$\tilde{\Sigma} = \begin{pmatrix} 1 & \rho_{23} \\ \rho_{23} & 1 \end{pmatrix} - \frac{1}{\sigma_1^2} \begin{pmatrix} \rho_{12}^2\sigma_1^2 & \rho_{12}\rho_{13}\sigma_1^2 \\ \rho_{12}\rho_{13}\sigma_1^2 & \rho_{13}^2\sigma_1^2 \end{pmatrix},$$

which are derived from standard properties of the multivariate normal distribution (see, for example, [28]). For notational convenience we drop the conditioning on d_i and y_{i0}.

Since Z_{i2} and Z_{i3} are not observed directly, the contribution of patient i to the likelihood is:

$$f(Z_{i1}; \theta) \int_{A(F_i)} \int_{A(T_i)} g_2(x, \tilde{\mu}, \tilde{\Sigma}) dx, \quad (5)$$

where $A(u) = (0, \infty)$ if $u = 1$ and $(-\infty, 0)$ if $u = 0$. Here, $g_K(x, m, S)$ denotes the pdf of a K-dimension normal distribution with mean m and covariance matrix S, evaluated at x.

The likelihood function is maximised to find the maximum likelihood estimator (MLE) of θ, $\hat{\theta}$. To do this in an efficient manner, the integrals in (5) are evaluated using the method of Genz and Bretz[29], which is a highly efficient Monte-Carlo integration method

that can be used for integrating over the multivariate normal or t-distribution. The overall optimisation is done using optim in R [30] with the "BFGS" option used. The covariance matrix of the MLE, is estimated as $\widehat{\text{Cov}}(\hat{\theta})$, the inverse of the Hessian matrix. To ease estimation, we introduce parameters $\delta = (\delta_1, \delta_{12}, \delta_{13}, \delta_{23})$ such that:

$$\begin{aligned} \sigma_1^2 &= \exp(\delta_1) \\ \rho_{jk} &= 2\frac{\exp(\delta_{jk})}{1+\exp(\delta_{jk})} - 1 \ (j,k) \in \{(1,2),(1,3),(2,3)\}, \end{aligned} \tag{6}$$

These reparameterisations induce the required range for the variance and correlation parameters while allowing estimation of parameters that can take any real value.

Once $\hat{\theta}$ and $\widehat{\text{Cov}}(\hat{\theta})$ are calculated, we can find the predicted probability of various combinations of toxicity and efficacy given a certain dose. Table 1 shows the various events that may be of interest in a dose-finding trial, and how they can be expressed as integrals over the joint distribution of $Z_i = (Z_{i1}, Z_{i2}, Z_{i3})$.

For example, the probability of efficacy without toxicity in individual i with baseline tumour size y_{i0} and dose d_i, $p_{E\bar{T}}$, is:

$$p_{E\bar{T}}(d_i, y_{i0}) = \int_{-\infty}^{0}\int_{-\infty}^{0}\int_{-\infty}^{\log(0.7)} g_3\left((z_{i1}, z_{i2}, z_{i3}), \bar{\mu}(d_i, y_{i0}), \bar{\Sigma}\right) dz_{i1}dz_{i2}dz_{i3},$$

where $\bar{\mu}(d_i, y_{i0}) = \begin{pmatrix} \mu_1 + \alpha_1 d_i + \beta_1 y_{i0} \\ \mu_2 + \alpha_2 d_i + \beta_2 y_{i0} \\ \mu_3 + \alpha_3 d_i + \beta_3 y_{i0} \end{pmatrix}$, and $\bar{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1 & \rho_{13}\sigma_1 \\ \rho_{12}\sigma_1 & 1 & \rho_{23} \\ \rho_{13}\sigma_1 & \rho_{23} & 1 \end{pmatrix}$. The upper limit of $\log(0.7)$ comes from response being defined by a greater than 30% shrinkage from baseline, equivalent to the log tumour size ratio being less than

$\log(0.7)$. The estimated probability of $p_{E\bar{T}}(d, y_{i0})$, $\hat{p}_{E\bar{T}}(d, y_{i0})$, has the same form but with all true parameter values replaced by their MLEs.

We are typically interested in the average probability of efficacy without toxicity across all individuals with baseline tumour sizes $y_0 = (y_{10}, \ldots, y_{N0})$ were they given dose d, defined as $\pi_{E\bar{T}}(d) = \frac{\sum_{i=1}^{N} p_{E\bar{T}}(d, y_{i0})}{N}$, with its estimate $\hat{\pi}_{E\bar{T}}(d) = \frac{\sum_{i=1}^{N} \hat{p}_{E\bar{T}}(d, y_{i0})}{N}$.

To estimate the variance of $\hat{\pi}_{E\bar{T}}(d)$, we can use the delta method:

$$\mathrm{Var}(\hat{\pi}_{E\bar{T}}) \approx (\bigtriangledown \hat{\pi}_{E\bar{T}})^T \widehat{\mathrm{Cov}}(\hat{\theta})(\bigtriangledown \hat{\pi}_{E\bar{T}}), \tag{7}$$

where $\bigtriangledown \hat{\pi}_{E\bar{T}}$ is the vector of partial derivatives of $\hat{\pi}_{E\bar{T}}$ with respect to each entry of θ. These are evaluated numerically. The estimated variance in equation (7) can be used to form a 95% confidence interval (CI) for $\pi_{E\bar{T}}$:

$$\left(\hat{\pi}_{E\bar{T}}(d) - \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\sqrt{\mathrm{Var}(\hat{\pi}_{E\bar{T}}(d))}, \hat{\pi}_{E\bar{T}}(d) + \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\sqrt{\mathrm{Var}(\hat{\pi}_{E\bar{T}}(d))}\right). \tag{8}$$

As shown in Wason and Seaman[24], the coverage of the CI may be improved if one works on the log-odds scale instead of the probability scale, especially if the average probability is near to 0 or 1.

After the model has been fitted, this process can be repeated for a range of doses to get a plot of the relationship between d and the quantity of interest, e.g. $\hat{\pi}_{E\bar{T}}(d)$. Both the estimated quantity and confidence bands are yielded from this procedure.

We henceforth refer to this approach as the augmented binary method, although in fact it is an extension of the original augmented binary method to allow for toxicity and dose dependence.

Binary methods

We compare the proposed method to two methods that use the dichotomised efficacy outcome only. The first, referred to as binary 1,

is, for each dose given to patients, to estimate the quantity of interest as the observed proportion together with 95% CI using a Wilson interval (binconf in the Hmisc package of R). This method does not attempt to model the relationship between dose and efficacy/toxicity and in particular cannot estimate a probability of efficacy without toxicity for a dose not given.

The second method, binary 2, fits a logistic regression to the binary indicator of the outcome of interest with parameters for baseline tumour size, dose and dose squared. We include linear and quadratic terms for dose to allow flexibility in the dependence of dose on probability of efficaccy without toxicity, however alternative models, such as splines, could be used if the sample size allows. The model is:

$$\log\left(\frac{p}{1-p}\right) = \mu + \alpha_1 d_i + \alpha_2 d_i^2 + \beta_1 y_{i0}, \tag{9}$$

where p is the probability of the outcome of interest.

From the model in equation (9), the parameter estimates are used to get predicted probability of the quantity of interest for each individual in the study had they been given dose d. The mean of this quantity (e.g. $\pi_{E\bar{T}}(d)$ when the outcome of interest is efficacy without toxicity), and a CI estimated from the delta method, can be found for different values of d to form a dose-response curve with 95% confidence bands.

Simulation studies

To understand the operating characteristics of the three methods, we performed a set of simulation studies. In all cases we assume five doses ($d \in (0, 1, 2, 3, 4)$) are available with an equal number of patients allocated to each. Although this is a larger number than the three doses tested in the case study (described later), we felt this would better represent potential wider applications to early-phase trials.

Simulation study methods

In the first simulation study, we explored the properties of the augmented binary method when the assumptions made by the model are broadly true. In all cases the baseline tumour size, y_{i0}, is simulated as a Uniform(1,10) random variable. Given dose, d_i, and y_{i0}, the log tumour size ratio Z_{i1} is assumed to follow a normal distribution with variance 1 and mean equal to an intercept and a linear effect of dose. The log odds of toxicity and new lesions also (separately) depend linearly on dose. Mathematically:

$$\begin{aligned} Z_{i1} &= \mu_1 + \alpha_1 d_i + \varepsilon_{i1} \\ \mathbb{P}(F_i = 1 | d_i, Z_{i1}) &= \text{expit}(\mu_2 + \alpha_2 d_i) \\ \mathbb{P}(T_i = 1 | d_i) &= \text{expit}(\mu_3 + \alpha_3 d_i), \end{aligned} \tag{10}$$

where $\varepsilon_{i1} \sim N(0, \sigma^2)$ is independent of d_i and $\text{expit}(x) = \frac{\exp(x)}{1+\exp(x)}$. In this case, conditional on dose, the three variables are independent. The values for $\mu_1, \alpha_1, \mu_2, \alpha_2, \mu_3, \alpha_3$ are chosen to form four different scenarios. Table 2 shows the parameter values used in the four scenarios. We note that the data-generating model (10) uses the logistic link whereas the analysis model uses the probit link. This is the case so that sensitivity of results to the simulation assumptions being different to the augmented model assumptions could be assessed.

Scenario 1 represents no effect of dose on any of the outcomes. Scenario 2 represents a cytotoxic drug where increasing dose leads to better tumour shrinkage and a reduced chance of new lesions but higher toxicity. Scenario 3 represents a cytostatic drug that reduces the probability of new lesions but does not affect tumour size or toxicity. Scenario 4 represents an ineffective and toxic drug where increased dose leads to higher toxicity but no change in efficacy.

In all cases we assumed that 15 patients had been treated with each of five doses $d = (0, 1, 2, 3, 4)$, thus 75 patients in total. For scenario 1 we also considered having 6 patients per dose and 25 patients per dose, to determine if there was a noticeable effect of the sample size on the properties of the method.

We estimated the statistical properties of the three methods using the true dose-response relationship yielded from the underlying model. In particular we estimated the coverage, the width of the confidence interval, the bias and mean squared error for each method; we also estimated the width of the confidence band and mean absolute error (i.e. the expected absolute difference between predicted and true values) across the range of doses for the binary 2 and augmented binary method. These latter two quantities were found for a range of doses between 0 and 4, with the trapezium rule used to approximate the integral of the quantity between d=0 and 4.

The second simulation study explored the situation where the probability of toxicity was associated with the log tumour size ratio (inducing a correlation). This was done by simulating the data using the following equations:

$$\begin{aligned} Z_{i1} &= \mu_1 + \alpha_1 d_i + \varepsilon_{i1} \\ \mathbb{P}(F_i = 1 | d_i, Z_{i1}) &= \text{expit}(\mu_2 + \alpha_2 d_i) \\ \mathbb{P}(T_i = 1 | d_i, Z_{i1}) &= \text{expit}(\mu_3 + \alpha_3 d_i + \beta_1 Z_{i1} + \beta_2 Z_{i1}^2). \end{aligned} \tag{11}$$

By changing the value of (β_1, β_2), the extent of the correlation between Z_{i1} and T_i can be varied. We note that the form of this correlation is not necessarily consistent with the form allowed for in the latent normal model assumed in the augmented binary analysis method. Thus, this setup should test whether the results from this method are sensitive to misspecification of the correlation form.

The third simulation study examined the properties of the augmented binary method whendata were simulated using a form for the probability of toxicity and mean tumour shrinkage that deviated strongly from the assumed latent variable model. One simulation scenario used the Emax model (see e.g. Macdougall[31]) for the toxicity probability; a second used the Emax model for the mean log-tumour size ratio. The respective models used to simulate data were:

$$\begin{aligned}\mathbb{P}(T_i = 1|d_i, y_{i0}, Z_{i1}) &= E_0 + \frac{(d_i^\lambda E_{\text{Max}})}{(d_i^\lambda + \text{ED50})},\\ \mathbb{E}(Z_i|d_i, y_{i0}) &= E_0 + \frac{(d_i^\lambda E_{\text{Max}})}{(d_i^\lambda + \text{ED50})}\end{aligned} \tag{12}$$

where E_0 represents the probability of toxicity or mean log tumour ratio when $d_i = 0$, $E_{\max}$ is the maximum effect attributable to the drug, ED50 is the dose that gives half the maximum effect, and λ determines the slope of the curve. In each case, other components were simulated as in Scenario 1. For the toxicity simulation study, the value of $(E_0, E_{\text{Max}}, \text{ED50}, \lambda)$ was set to $(0.1, 0.7, 2, 4)$. For the mean tumour shrinkage simulation study $(E_0, E_{\text{Max}}, \text{ED50}, \lambda)$ was set to $(0, \log(0.5), 2, 4)$. We investigated fitting the model given in equation (2) and the same model but with an additional parameter in the relevant part of model to allow (respectively) the mean of Z_{i1} or Z_{i3} to depend on the squared-dose. In all simulation studies, 5000 replicates were used.

Simulation study results

We first present the results of the first simulation study investigating four different scenarios (described in table 2). In figure 1, we show the true relationship between dose and probability of efficacy

without toxicity. Supplementary figure 1 shows the probability of the individual components (toxicity, tumour shrinkage and new lesions).

In supplementary table 1, the coverage of the two binary methods and the augmented binary method is provided at the five doses considered. Each of the methods appears to have around 95% coverage in most cases. In some situations the coverage appears to be higher or lower than the nominal level, even allowing for Monte-Carlo standard error (with 5000 replicates, this is 0.003). There are no cases where the coverage is worryingly low (the minimum coverage for the augmented binary method is 0.938).

In case the properties of the method were sensitive to the value of σ assumed in the simulations, we considered increased values of σ. Supplementary table 2 shows that this made no difference to the coverage. We also did a simulation scenario where the number of patients allocated to dose 2 was three times the number allocated to other doses - the coverage was unaffected by this also.

Table 3 provides the precision, in terms of the confidence interval widths, for the three methods in the four scenarios. We present the average confidence interval width for dose $d = 1$ and the overall area within the upper and lower confidence bands between doses 0 and 4 (as found by first calculating CIs for a grid of 100 dose values between $d = 0$ and $d = 4$ and using the trapezium rule to approximate the area between the confidence bands in the region $[0, 4]$). The binary 1 method does not give confidence bands, so only the average CI width for dose 1 is presented for it. The results show that binary 2 improves the precision at dose 1 compared to binary 1. This indicates that using a suitable model to borrow strength between data on different doses improves the precision at each dose. The results show that making use of the continuous tumour information further improves the efficiency: both the confidence interval width at dose 1 and the

area of the confidence band are considerably narrower using the augmented binary method compared to the binary 2 method.

The bias, mean squared error and expected absolute error are provided in supplementary tables 3-5. The bias of all methods is small for all doses. The augmented binary method provides a clear advantage in the mean squared error and mean absolute error in all scenarios.

We repeated scenario 1 with 6 patients per arm and 25 patients per arm to check how the relative performance of the methods may vary as the sample size changes. These results are presented in table 3 and supplementary table 1. The augmented binary methods coverage appears largely unaffected. The binary 2 method appears to become more conservative for smaller sample sizes. The relative performance of the methods appears similar for the different sample sizes considered. We note that the width of the CI found from the Aug Bin method at $d = 1$ for $n = 15$ per dose is slightly narrower than the width of the Binary 2 method at $d = 1$ for $n = 25$.

To investigate the effect of correlation between the different components of the outcome, we investigated varying β_1 in equation (11), whilst keeping $\beta_2 = 0$. Supplementary figure 2 shows that the coverage of the binary 2 and augmented binary methods remain similar as β_1 changes, perhaps decreasing slightly. The binary 1 method has coverage that changes sharply without a clear trend. This is likely due to the discrete nature of the binomial distribution, which may affect binary 1 (which does not borrow information over doses) more severely than it affects the other two methods. We also investigated the coverage of confidence intervals from the method if the latent variable model that assumed independence between the components This was done by fitting the latent variable model with correlation parameters constrained to be 0. For positive values of β_1,

this led to substantial bias and lower than nominal coverage when estimating the probability of efficacy without toxicity. For $\beta_1 = 1$, the coverage at particular doses was as low as 80%. This indicates that modelling the correlation is important.

Sensitivity to assumptions

The previous simulation scenarios have been approximately consistent with the augmented binary model's assumptions. We conducted further simulation studies to assess how the method might perform in scenarios where the model assumptions are further from holding.

We first examined the properties of the augmented binary method when a non-linear relationship between the dose and log-odds of toxicity was present. This was done by simulating toxicity data using the Emax model in equation (12), with all other parameters set as in scenario 1. In this case the dose has the effect of increasing toxicity but not efficacy. The augmented binary method using the models in (2) was fitted. In addition, the method was applied using a modified model that also included a parameter for the effect of the squared dose (d_i^2) on the toxicity latent variable Z_{i3}. We repeated this but with an emax model used to simulate the mean log tumour size ratio. The coverage results for both models in both scenarios are provided in supplementary table 6. We see that the coverage of the augmented binary method is highly dependent on being able to model the true relationship between dose and the respective component reasonably well. The model described by (2) leads to very poor coverage for some doses, especially $d = 2$, which is where the ED50 is. The modified model that includes a quadratic term does much better, with the worst coverage being 92.6% at $d = 2$ for the toxicity scenario.

As a second sensitivity analysis we examined what happened if a non-zero value of β_2 in (11) was used. This represents a

quadratic effect of the log tumour ratio on the probability of toxicity. Supplementary figure 3 shows that the coverage appears to be slightly lower than nominal when β_2 is negative. For positive values of β_2, coverage is correct or slightly higher than nominal.

Case study

The HORIZON II trial initially randomised patients $1:1:1$ between arms. However at an interim analysis the 30 mg dose was closed to further recruitment and subsequently recruited patients were randomised $2:1$ between 20mg and placebo. In total, the numbers of patients randomised to placebo, 20mg and 30mg were 346, 484 and 209 respectively.

We applied the two binary methods, the augmented binary model described in equation (2) (referred to as the AugBin linear dose model) and a slightly more complicated augmented binary model that added a quadratic dose term for each of the three components (referred to as the AugBin quadratic dose model):

$$\begin{aligned} Z_{i1} &= \mu_1 + \alpha_1 d_i + \beta_1 d_i^2 + \gamma_1 y_{i0} + \varepsilon_{i1} \\ Z_{i2} &= \mu_2 + \alpha_2 d_i + \beta_2 d_i^2 + \gamma_2 y_{i0} + \varepsilon_{i2} \\ Z_{i3} &= \mu_3 + \alpha_3 d_i + \beta_3 d_i^2 + \gamma_3 y_{i0} + \varepsilon_{i3}, \end{aligned} \tag{13}$$

The dose scale was standardised so that the value of d used in the models was 0 for placebo, 1 for 20mg and 2 for 30mg.

We first explored the best transformation to apply to the tumour size change by finding the best Box-Cox transformation. This indicated that the tumour ratio (Y_{i1}/Y_{i0}) was closer to normally distributed than the log tumour ratio Z_{i1}, hence we used the former as the outcome of the continuous part of the augmented binary models in (2). The probability of efficacy without toxicity was estimated

for a grid of standardised doses in the region $[0, 2]$ for the binary 2 method and the two augmented binary methods. The estimated curves, together with confidence bands, are plotted in figure 2. Table 4 reports the estimated probability of efficacy without toxicity for each of the three doses together with the confidence band area for the binary 2 and augmented binary methods.

Both augmented binary methods give substantial reductions in the confidence interval widths and the area between the upper and lower confidence bands compared to the binary methods (table 4). Although it is difficult to determine whether changes to the dose-response relationship are due to chance, it appears that including a quadratic term changes the conclusion as to which is the superior dose including it indicates that 20mg is the dose that best balances safety and efficacy. When measured by AIC the quadratic model fits slightly better (difference in AIC of around 1). These results indicate that if the sample size permits, investigating a range of possible models may be beneficial.

Discussion

In this paper we have considered trials conducted to explore the efficacy and toxicity of different doses of a novel treatment. Many treatments fail later in the drug-development process due to lack of efficacy or unacceptable toxicity [32, 33], so it is important to ensure the right dose is chosen to balance these two factors. Methods that can improve the chance of picking the most suitable dose are therefore very important.

We have adapted a method [24] that was originally proposed to improve the efficiency of analyses when the efficacy endpoint is composite with one continuous component and one binary component. This is the case for phase II oncology trials where a patient

is a responder if their target tumour lesions shrink by at least 30% and they do not have any new lesions or other reasons for treatment failure (such as non-target lesions increasing unacceptably in size). That method involved fitting a suitable joint model to the different components and then drawing inferences about the composite endpoint. This had the property of better utilising the continuous component, for example by weighting observed responses less if the patients tumour shrinkage was close to the 30% threshold than if their shrinkage was much greater than the threshold.

Here, we have adapted that method to allow for a relationship between dose and the probabilities of toxicity and efficacy. This was done through a multivariate probit model, where the two binary components (toxicity and new lesions) are included using latent variables. This model allows better modelling of correlation between the two efficacy and one toxicity components than previous methods allowed. This is particularly important in phase I/II trials, because for many types of cancer treatment there is likely to be a correlation between efficacy and toxicity. Although not considered, the method allows additional covariates to be adjusted for also.

Our simulation study demonstrated that this method can result in a substantial gain in efficiency compared to methods that do not exploit the continuous nature of the tumour size, and in the HORIZON II study the width of the confidence interval for the probability of efficacy without toxicity was considerably narrower. The simulation study also showed that the coverage of the confidence interval is generally close to nominal if model assumptions are correct. We also assessed the sensitivity of coverage to some of the model assumptions. It was clear from these investigations that it is important that the models used for each of the components allow the modelled relationship to be close to the true one. For example, when

a non-linear relationship between dose and probability of toxicity was present, the model performed much better when both linear and quadratic effects of dose on toxicity were allowed compared to when just a linear effect was allowed for. The coverage of the model appeared to remain good when more complicated correlations between the components of the outcome were simulated.

Several important considerations emerged from the case study. First it is important to have a good understanding, prior to starting the trial, of the likely impact of dose on the various components so that a suitable model can be pre-specified. Different conclusions resulted from the augmented binary model with just a linear dose term compared to one with linear and quadratic terms. Generally models that allow for more flexiblility in the shape of the dose-response relationship should be preferred if the sample size is sufficiently high. Second, it is important to transform the tumour shrinkage outcome appropriately so that the residuals of the model are as close to normally distributed as possible. The results from the augmented binary method were quite different when the log tumour ratio was used as opposed to the Box-Cox transform which resulted in residuals being closest to normality. Third, there were a small number of complete responses observed (100% tumour shrinkage) which means that the assumption of normally distributed outcomes is not met. Work to apply a more suitable distribution such as the censored normal distribution [34] might be of use if there is a high proportion of complete responses. Work is underway to apply the method to a few different case studies and propose guidelines on a workflow for applying the method in a trial.

Our results have also concentrated on inferring the dose curve rather than testing the difference in probability of response without toxicity of doses against that of control. In phase IIb trials it is often

of interest to formally test doses against a placebo or control in terms of efficacy. This would be possible to do using our proposed method by assuming that placebo/control can be considered as a dose of 0. Approaches such as MCP-mod [35] have been proposed to simultaneously allow for uncertainty in the underlying dose-response curve whilst testing doses against control. This would be very complementary to the methodology proposed here; investigating how the two might be combined would be both interesting and help address issues mentioned in the previous paragraph.

In this paper we have focused on estimating the probability of efficacy and no toxicity. Other quantities can be estimated using the method, including the probability of efficacy given a toxicity constraint and a weighted score of probability and toxicity. In fact, one benefit of the method is that several of these quantities can be estimated simultaneously and their joint distribution be found. This would allow for more efficient control of the type I error rate if multiple testing correction is required.

We have focused on oncology trials in this paper, motivated by the HORIZON II dataset. However, the method may be applicable in several other areas. Trial endpoints that are a composite of continuous and binary components are used in several other areas. One example is rheumatoid arthritis [36], where several disease scoring criteria have this property. It is important to bear in mind that sample sizes in different disease areas may vary and that the HORIZON II trial was much larger than would be expected for a typical trial assessing efficacy and safety of different doses. Previous work by McMenamin et al [37] has examined the statistical properties of the augmented binary method for small sample size situations and has proposed corrections to improve the properties. It would be of interest to extend this work to the latent variable model used in this paper. On the

other hand, for trials with large sample sizes it would be of interest to explore how more flexible modelling approaches such as splines could be applicable.

There are several extensions to this work which would be of interest. The first is to apply the method in adaptive phase I/II trials to help select the next dose. The main issue here would be applying the method to small samples; this would be especially true for early interim analyses. It would be possible to use an alternative approach early on in the trial and switch to using the augmented method when there are sufficient data available. In this case it would be of interest to compare the resulting method to current adaptive methods [3, 4, 5, 6, 7, 8, 9, 11, 12]. Since these methods either analyse response as binary or categorical (which means they are likely to have less power), or analyse it as a continuous variable only (which means they are likely to have less clinical interpretability for RECIST-based response), the new method is potentially very useful.

A second extension would be to consider more than one follow-up time. By including more components in the latent variable model, it would be possible to allow for the efficacy or toxicity endpoint to be observed at several timepoints. Although possible, having more timepoints would increase the number of parameters, particularly ones representing the covariance. It would also be more computationally complex to maximise the likelihood and perform the multivariate integration to infer the probability of efficacy without toxicity. Thus, careful consideration of the correlation structure might be necessary to avoid having too many parameters or an infeasible computational burden. Despite these difficulties, it would be worthwhile to pursue this in order to allow use of the methodology for a wider variety of phase I/II trials. Some related work on reducing the computational

burden in the context of multiple efficacy timepoints is presented in Lin and Wason [38].

A third extension would be to develop the method to use more detailed information on toxicity. We assumed toxicity was represented by a binary component, but in fact it is potentially more complex than this. We could extend the latent variable model to allow for an ordinal toxicity variable (bringing in grading of observed toxicities). Going further, a model could be developed that uses the richness of toxicity data gathered in practice to the fullest extent possible. Toxicity criteria such as the Common Toxicity Criteria [18] contain many different components, some of which are based on continuous values. The challenge would be to fit a suitable model to such a high number of components, although the resulting efficiency gains may be even higher than those seen in this work.

In conclusion, we have proposed a method that can substantially increase the efficiency of trials that assess different doses of a treatment on both a toxicity and composite efficacy endpoint.

Software

An R package, AugBinDoseFinding is available from https://sites.google.com/site/jmswason/supplementary-material.

Acknowledgements

This work was supported by the Medical Research Council (grant numbers `MC_UP_1302/4` and `MC_UU_00002/10`), Cancer Research UK (grant number C48553/A18113). The authors thank AstraZeneca for permission to use the HORIZON data.

References

[1] Iasonos A and OQuigley J. Early phase clinical trials – are dose expansion cohorts needed? *Nature reviews Clinical oncology* 2015; 12(11): 626.

[2] Yin G. *Clinical trial design: Bayesian and frequentist adaptive methods*, volume 876. John Wiley & Sons, 2013.

[3] Bryant J and Day R. Incorporating toxicity considerations into the design of two-stage phase ii clinical trials. *Biometrics* 1995; 51(4): 1372–1383.

[4] Braun TM. The bivariate continual reassessment method: extending the CRM to phase I trials of two competing outcomes. *Controlled clinical trials* 2002; 23(3): 240–256.

[5] Thall PF and Cook JD. Dose-finding based on efficacy–toxicity trade-offs. *Biometrics* 2004; 60(3): 684–693.

[6] Nebiyou Bekele B and Shen Y. A Bayesian approach to jointly modeling toxicity and biomarker expression in a phase I/II dose-finding trial. *Biometrics* 2005; 61(2): 343–354.

[7] Zhang W, Sargent DJ and Mandrekar S. An adaptive dose-finding design incorporating both toxicity and efficacy. *Statistics in medicine* 2006; 25(14): 2365–2383.

[8] Yin G, Li Y and Ji Y. Bayesian dose-finding in phase I/II clinical trials using toxicity and efficacy odds ratios. *Biometrics* 2006; 62(3): 777–787.

[9] Dragalin V, Fedorov VV and Wu Y. Two-stage design for dose-finding that accounts for both efficacy and safety. *Statistics in Medicine* 2008; 27(25): 5156–5176.

[10] Zhong W, Koopmeiners JS and Carlin BP. A trivariate continual reassessment method for phase i/ii trials of toxicity, efficacy, and surrogate efficacy. *Statistics in medicine* 2012; 31(29): 3885–3895.

[11] Wages NA and Tait C. Seamless phase I/II adaptive design for oncology trials of molecularly targeted agents. *Journal of biopharmaceutical statistics* 2015; 25(5): 903–920.

[12] Thall PF and Nguyen HQ. Adaptive randomization to improve utility-based dose-finding with bivariate ordinal outcomes. *Journal of biopharmaceutical statistics* 2012; 22(4): 785–801.

[13] Liu S and Johnson VE. A robust bayesian dose-finding design for phase i/ii clinical trials. *Biostatistics* 2016; 17(2): 249–263.

[14] Yeung WY, Reigner B, Beyer U et al. Bayesian adaptive dose-escalation designs for simultaneously estimating the optimal and maximum safe dose based on safety and efficacy. *Pharmaceutical statistics* 2017; 16(6): 396–413.

[15] Mozgunov P and Jaki T. An information theoretic phase I–II design for molecularly targeted agents that does not require an assumption of monotonicity. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 2018; .

[16] Yuan Y, Nguyen HQ and Thall PF. *Bayesian designs for phase I–II clinical trials*. Chapman and Hall/CRC, 2016.

[17] Eisenhauer E, Therasse P, Bogaerts J et al. New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). *European Journal of Cancer* 2009; 45: 228–247.

[18] Common terminology criteria for adverse events (CTCAE) v5.0, 2018. URL `https://ctep.cancer.gov/protocoldevelopment/`

electronic_applications/docs/CTCAE_v5_Quick_Reference_8.5x11.pdf.

[19] Thall PF, Nguyen HQ and Zinner RG. Parametric dose standardization for optimizing two-agent combinations in a phase I–II trial with ordinal outcomes. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 2017; 66(1): 201–224.

[20] Lee J, Thall PF, Ji Y et al. A decision-theoretic phase I–II design for ordinal outcomes in two cycles. *Biostatistics* 2016; 17(2): 304–319.

[21] Houede N, Thall PF, Nguyen H et al. Utility-based optimization of combination therapy using ordinal toxicity and efficacy in phase I/II trials. *Biometrics* 2010; 66(2): 532–540.

[22] Fedorov V, Wu Y and Zhang R. Optimal dose-finding designs with correlated continuous and discrete responses. *Statistics in medicine* 2012; 31(3): 217–234.

[23] Hirakawa A. An adaptive dose-finding approach for correlated bivariate binary and continuous outcomes in phase I oncology trials. *Statistics in medicine* 2012; 31(6): 516–532.

[24] Wason J and Seaman SR. Using continuous data on tumour measurements to improve inference in phase II cancer studies. *Statistics in medicine* 2013; 32(26): 4639–4650.

[25] Hoff PM, Hochhaus A, Pestalozzi BC et al. Cediranib plus FOLFOX/CAPOX versus placebo plus FOLFOX/CAPOX in patients with previously untreated metastatic colorectal cancer: a randomized, double-blind, phase III study (HORIZON II). *Journal of Clinical Oncology* 2012; 30(29): 3596–3603.

[26] Lavin P. An alternative model for the evaluation of antitumor activity. *Cancer Clinical Trials* 1981; 4: 451–457.

[27] Chib S and Greenberg E. Analysis of multivariate probit models. *Biometrika* 1998; 85(2): 347–361.

[28] Eaton ML. Multivariate statistics: a vector space approach. *JOHN WILEY & SONS, INC, 605 THIRD AVE, NEW YORK, NY 10158, USA, 1983, 512* 1983; .

[29] Genz A and Bretz F. Methods for the computation of multivariate t-probabilities. *Journal of Computational and Graphical Statistics* 2002; 11: 950–971.

[30] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017. URL `https://www.R-project.org/`.

[31] Macdougall J. *Analysis of Dose–Response Studies—Emax Model*, chapter 9. New York, NY: Springer New York. ISBN 978-0-387-33706-7, 2006. pp. 127–145. DOI:10.1007/0-387-33706-7_9. URL `https://doi.org/10.1007/0-387-33706-7_9`.

[32] Kola I and Landis J. Can the pharmaceutical industry reduce attrition rates? *Nature Reviews Drug Discovery* 2004; 3: 711–716.

[33] Seruga B, Ocana A, Amir E et al. Failures in phase III: causes and consequences, 2015.

[34] Wason JM and Mander AP. The choice of test in phase II cancer trials assessing continuous tumour shrinkage when complete responses are expected. *Statistical methods in medical research* 2015; 24(6): 909–919.

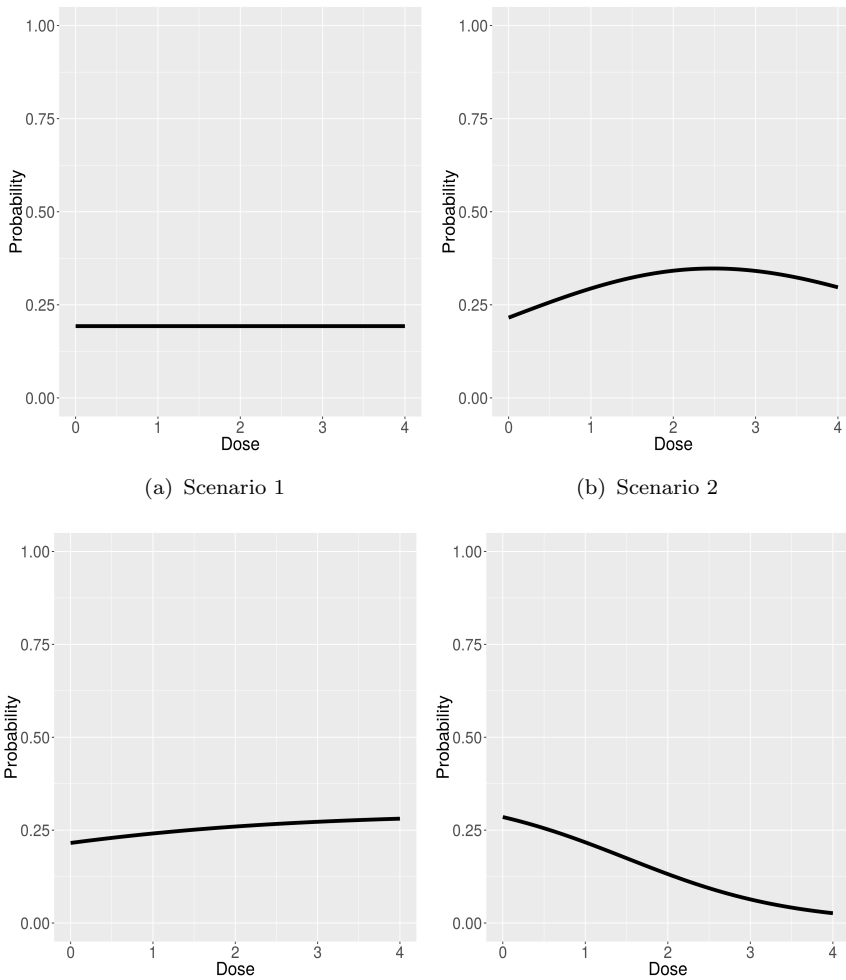
[35] Bornkamp B, Pinheiro J, Bretz F et al. MCPMod: An R package for the design and analysis of dose-finding studies. *Journal of Statistical Software* 2009; 29(7): 1–23.

[36] Wason JM and Jenkins M. Improving the power of clinical trials of rheumatoid arthritis by using data on continuous scales when analysing response rates: an application of the augmented binary method. *Rheumatology* 2016; 55(10): 1796–1802.

[37] McMenamin M, Berglind A and Wason JM. Improving the analysis of composite endpoints in rare disease trials. *Orphanet journal of rare diseases* 2018; 13(1): 81.

[38] Lin CJ and Wason JM. Improving phase II oncology trials using best observed recist response as an endpoint by modelling continuous tumour measurements. *Statistics in medicine* 2017; 36(29): 4616–4626.

Figure 1. Actual relationship between dose and probability of efficacy without toxicity in simulation scenarios.



(a) Scenario 1

(b) Scenario 2

(c) Scenario 3

(d) Scenario 4

Figure 2. Plot of relationship between standardised dose and probability of efficacy without toxicity for the binary method and two augmented binary methods. Solid lines indicate the estimated curve and dotted lines represent the 95% confidence bands

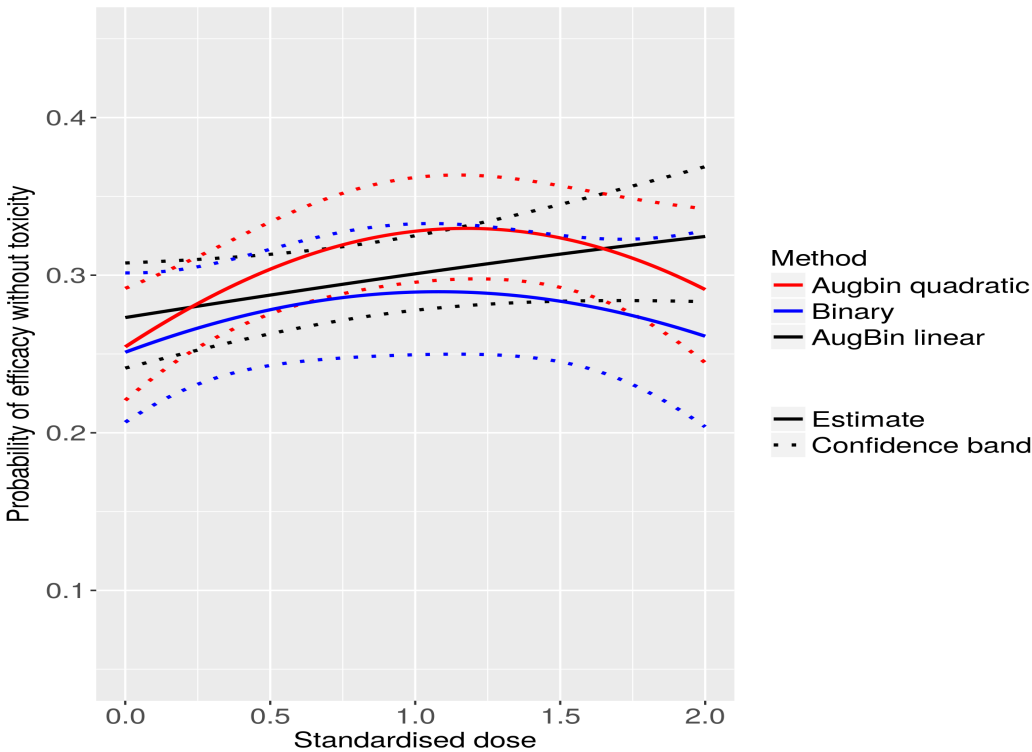


Table 1. Summary of possible events of interest to model in dose-finding trial

Outcome for patient	Required events for this outcome	Modelled probability of event in terms of latent variable model
Efficacy	$F_i = 0$ and $z_{i1} < \log(0.7)$	$\int_{-\infty}^{0}\int_{-\infty}^{\log(0.7)} f_{Z_{i1},Z_{i2}}(z_1,z_2)dz_1dz_2$
Toxicity	$T_i = 1$	$\int_{0}^{\infty} f_{Z_{i3}}(z_3)dz_3$
Efficacy without toxicity	$F_i = 0, z_{i1} < \log(0.7)$ and $T_i = 0$	$\int_{-\infty}^{0}\int_{-\infty}^{0}\int_{-\infty}^{\log(0.7)} f_{(Z_{i1},Z_{i2},Z_{i3})}(z_1,z_2,z_3)dz_1dz_2dz_3$
Efficacy with toxicity	$F_i = 0, z_{i1} < \log(0.7)$ and $T_i = 1$	$\int_{0}^{\infty}\int_{-\infty}^{0}\int_{-\infty}^{\log(0.7)} f_{(Z_{i1},Z_{i2},Z_{i3})}(z_1,z_2,z_3)dz_1dz_2dz_3$
Toxicity without efficacy	$T_i = 1$ and at least one of $F_i = 1$ or $z_{i1} > \log(0.7)$	$\int_{0}^{\infty} f_{Z_{i3}}(z_3)dz_3 -$ $\int_{0}^{\infty}\int_{-\infty}^{0}\int_{-\infty}^{\log(0.7)} f_{(Z_{i1},Z_{i2},Z_{i3})}(z_1,z_2,z_3)dz_1dz_2dz_3$
No efficacy and no toxicity	$T_i = 0$ and at least one of $F_i = 1$ or $z_{i1} > \log(0.7)$.	$\int_{-\infty}^{0} f_{Z_{i3}}(z_3)dz_3 -$ $\int_{-\infty}^{0}\int_{-\infty}^{0}\int_{-\infty}^{\log(0.7)} f_{(Z_{i1},Z_{i2},Z_{i3})}(z_1,z_2,z_3)dz_1dz_2dz_3$

Table 2. Description of the four scenarios used in simulation study 1.

Scenario	μ_1	α_1	μ_2	α_2	μ_3	α_3
1 - no effect of dose on toxicity or efficacy	0	0	-1	0	-1	0
2 - cytotoxic drug, shrinks tumour and increases toxicity	0	-0.4	-1	-0.25	-1.5	0.5
3 - cytostatic drug, reduces probability of new lesions but does not increase toxicity	0	0	-1	-0.5	-1.5	0
4 - ineffective and toxic drug	-0.3	0	-1	0	-1.5	1

Table 3. Efficiency results in terms of confidence interval width at dose 1 and area between the upper and lower confidence interval curves across all doses.

	Mean CI width for dose $d = 1$			Confidence band area	
	Binary 1	Binary 2	Aug Bin	Binary 2	Aug Bin
Scenario 1	0.364	0.246	0.167	1.088	0.689
Scenario 2	0.406	0.279	0.204	1.245	0.898
Scenario 3	0.388	0.260	0.185	1.177	0.776
Scenario 4	0.376	0.315	0.189	1.193	0.628
Scenario 1 $n = 6$ per dose	0.524	0.469	0.263	2.045	1.085
Scenario 1 $n = 25$ per dose	0.269	0.173	0.119	0.765	0.490

Table 4. Estimated probability of efficacy without toxicity, confidence interval, and area of confidence bands for the binary and augmented binary methods when applied to the HORIZON II dataset.

Method	Placebo	20mg	30mg	Area of confidence band
Binary 1	0.257 (0.213,0.306)	0.298 (0.259,0.341)	0.276 (0.218,0.343)	NA
Binary 2	0.251 (0.207,0.301)	0.289 (0.250,0.333)	0.261 (0.204,0.328)	0.167
Aug Bin linear	0.273 (0.241,0.308)	0.301 (0.278,0.325)	0.325 (0.283,0.369)	0.116
Aug Bin quadratic	0.254 (0.221,0.292)	0.328 (0.296,0.362)	0.291 (0.244,0.342)	0.132